Journal of Engg. Research Vol. 7 No. (3) September 2019 pp. 123-141

Intent Aware Optimization for Content-Based Lecture Video Retrieval Using Grey Wolf Optimizer

Sanjay B. Waykar* and Dr. C. R. Bharathi

**Research Scholar, Dept. of Computer Science & Engineering, Specialty of Image Processing*
Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, India
Associate Professor, Dept. of Electronics and Communications, Specialty of Image Processing
Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, India
**Corresponding Author: sanjaywaykarb@gmail.com*

ABSTRACT

Nowadays, video recordings are widely used and easy in spreading the knowledge among the students. Due to the rapid development of recording technologies and video-based learning, the large number of videos is published on the Internet. The main challenge is to retrieve the appropriate video based on the user requirement. This paper proposes the intent aware optimization based on grey wolf optimizer for retrieving the lecture video. The extraction of keyframe is the initial step in the proposed system. The next step is the key frame extraction in which the keywords from the key frame are recognized by the optical character recognition and LVP (Local Vector Pattern). After the features are extracted, the PENN (Probability Extended Nearest Neighbour) classifier is utilized to retrieve the relevant videos for the text or video query. Subsequently, the user selects one video, which is used for the matching purpose based on the optimization. The grey wolf optimizer is applied to the input database where the clustering task acquires the optimal solution. Finally, the user selected video is matched with the optimal solution to retrieve the more relevant video for the input query. The experimental results are validated, and the parameters used to analyze the performance are F-measure, Recall, and Precision. The performance is compared with the existing systems using MATLAB implementation. The higher precision value of 75% is attained by the proposed method, which ensures the efficient retrieval of content based lecture video.

Keywords: Video retrieval, character recognition, text and video query, PENN classifier, Grey wolf optimizer.

1. INTRODUCTION

Nowadays, the recording of lecture video is one of the emerging techniques on the web. The lecture video on the web can easily spread the knowledge throughout the world, and students acquire the knowledge based on their requirement. There are three kinds of modalities in the lecture video, which are a speech in the lecture video, graphics, and printed text in both video and image slides. The lecture video lies in two different modalities such as multi-modal and cross-modal (Nguyen *et al.,* 2014). The metadata extraction is an important aspect of the lecture video retrieval. Also, the keywords and topic-based segments are the two factors in the content based lecture video retrieval. The keywords are extracted from each frame by the character recognition method, which is annotated earlier in the lecture video (Li *et al.,* 2015; Balasubramanian *et al.,* 2015). The query is considered as the main concern of the lecture video retrieval, which is given by the user. The query is either to be text or video query. If it is text query, the videos, which contain the corresponding keywords, are retrieved. Similarly, the video query contains the keywords that are used to determine the relevant videos (Feng *et al.,* 2011).

The text-based video retrieval is commonly used in the earlier days. The text-based approach is based on the metadata, like video title, content definition, date, author, and comments about the video. The drawback of this approach is that the metadata did not quietly describe the semantic content of the video. Moreover, this approach acquires high-cost computation and inconsistency (Feng *et al.,* 2011). These issues are overcome by the Content-Based Video Retrieval

(CBVR) approach. In CBVR approach, the content is defined as the features of the lecture video like shape, texture, and color (Hernandez & Hernandez, 2014). The content-based video retrieval is the major concern for the users, which satisfies their queries regarding video content analysis. The CBVR includes three significant components: i) the user provides the text query or video query, which consists of video examples of the semantic concept; ii) the desired video is retrieved from the video database with regard to the user query; iii) based on the text or video query, the videos are sorted in the database according to the relevant data (Beltran & Pla, 2016).

The two major issues in the lecture video retrieval are, 1) how the feature information is extracted from the input video and 2) how the feature element is used to determine the similarity measure between two videos? The video retrieval system comprises two steps: They are i) Extraction of keyframe and ii) Feature extraction (Gao *et al.,* 2014; Hernandez & Hernandez, 2014). The key frame extraction is the initial step to retrieve its relevant video. In general, the video consists of some frames, and the frame is represented as the image. The important content of the data is detected using keyframe extraction (Va & Narayanana, 2015). Therefore, each keyframe is matched with the video recording segment. Then, the metadata is extracted from each keyframe based on the user query. The optical character recognition method is utilized to extract the data from its background (Yang *et al.,* 2011). Finally, the relevant video based on the user text or video query is obtained by the ranking framework (Aly *et al.,* 2013), vector space model, and probabilistic latent semantic analysis (Cheny *et al.,* 2014).

In this paper, we propose the new intent aware optimization using the grey wolf optimizer for the lecture video retrieval. Initially, the relevant videos based on the input query are obtained by the classifier. The input database and input query are fed into video retrieval system. The content-based video is retrieved by the three steps, which are key frame extraction, feature information, and PENN classifier. Normally, each video contains a large number of frames, and each frame is represented as an image. These frames are used to extort the key frame, which contains the significant information. Subsequently, the key frame is given as input to determine the feature information. Here, the optical character recognition and Local vector pattern are utilized to extract the visual or textual keywords. The classifier exploits these features to retrieve the relevant videos from the database. However, the appropriate video for the input query is retrieved from the resultant videos. The grey wolf optimization algorithm is used to determine the best solution iteratively. Finally, the user selected video is matched with the optimal solution and the matching process retrieves the content based lecture video significantly.

The main contribution of this paper is as follows:

- The grey wolf optimizer is utilized to design the intent aware optimization for the retrieval of content based lecture video.
- The new intent aware optimization is used to match the optimal solution with the user selected video to retrieve the more relevant video.

The paper is organized as follows: The review of the content based lecture video retrieval from eight research paper is described in section 2. Section 3 presents the problem description and challenges behind the approach. Section 4 briefly explains the intent aware optimization for the content based lecture video retrieval. The experimental results and performance are analyzed in section 5. Finally, section 6 concludes this paper.

2. LITERATURE REVIEW

Haojin Yang and Christoph Meinel (2014) presented an approach for lecture video retrieval based on the video and speech recognition system. Initially, the automatic video segmentation was applied, and keyframe was detected from the input video. Consequently, the metadata in each keyframe was determined by the optical character recognition and then the audio track about the video was recognized by the automatic speech recognition. The OCR and ASR (Automatic Speech Recognition) method were used for the keyword extraction in which the video and segment level keywords were employed to retrieve the content-based video. The experimental results and performance were analyzed, which proved the better retrieval performance.

Vidhya Balasubramanian *et al.* (2015) explained the multimodal metadata extraction in which the key phrases and topic-based segments were extracted for the lecture video retrieval. Here, the extraction process underwent by the audio transcripts and slide contents of the video. Finally, the hybrid approach composed of naive Bayes classifier and rule-based refiner was employed to define the metadata. The metadata extraction technique was evaluated using the different sources of videos. The results showed that this multimodal approach was effective in summarizing the lecture's content, potentially improving the user experience during retrieval and browsing.

Kaiyang Liao *et al.* (2014) demonstrated the sample based hierarchical adaptive K-means clustering method (SHAKM) for video retrieval. This algorithm exploited the multilevel sample strategy to handle the large databases. Subsequently, the k-means clustering was utilized to find the appropriate number of cluster and constructed the cluster tree. Also, the fast labeling scheme was used to assign each pattern in the dataset, which was close to the cluster. The experimental results were evaluated, and the very large datasets analyzed the performance. From the experimental results, it could be shown that the SHAKM method retrieved the lecture video efficiently and successfully.

Nhu Van Nguyen *et al.* (2015) described the multi-modal and cross-modal for the lecture video retrieval. The video contains the set of subjects, which was defined by the visual and textual words. It had two assumptions; the first assumption was that the video might contain multiple subjects and the second assumption was that the video document includes multiple modalities. The frame in the video was represented as the image. The visual and textual words in each slide were extracted by the text detection and graphical localization. The localization was computed by the frame sequence of the input video. Based on the different subjects, the lecture video was clustered into different groups. The results were calculated by the indexing and retrieval approach for the lecture video retrieval and enhanced the retrieval accuracy efficiently.

Ruben Fernandez Beltran and Filiberto Pla (2016) presented a Content-Based Video Retrieval approach to cope with the semantic gap challenge using latent topics. Firstly, a supervised topic model was employed to transform the classical retrieval approach into a class discovery problem. Subsequently, a probabilistic ranking function was deduced from that model to tackle the semantic gap between low level features and high level concepts. Finally, a short-term relevance feedback scheme was defined where queries can be initialized with samples from inside or outside the database. Several retrieval simulations had been carried out using three databases and seven different ranking functions to test the performance of the presented approach. Experiments revealed that the proposed ranking function was able to provide a competitive advantage within the content based retrieval field.

Huizhong Chen *et al.* (2014) described the Multi-modal Language Models (MLMs), which was used to adapt latent variable techniques for document analysis to exploring co-occurrence relationships in multi-modal data. The application of MLMs was to indexing text from slides and speech in lecture videos and subsequently employed a multi-modal probabilistic ranking function for lecture video retrieval. The MLM achieved highly competitive results against well-established retrieval methods such as the Vector Space Model and Probabilistic Latent Semantic Analysis. When noise is present in the data, retrieval performance with MLMs is enhanced with the quality of the spoken text extracted from the video.

Matthew Cooper (Yoo & Cho, 2007) determined the relative utility of automatically recovered text from these sources for lecture video retrieval. The slide was automatically detected within the videos and applied to the optical character recognition to obtain their text. Automatic speech recognition was used similarly to extract spoken text from the recorded audio. The experimentation was performed with manually created ground truth for both the slide and spoken text from more than 60 hours of lecture video. The automatically extracted slide and spoken text were compared regarding accuracy relative to ground truth, overlap with one another, and utility for video retrieval. Results showed that automatically recovered slide text and spoken text contain different content with varying error profiles. Experiments demonstrated that automatically extracted slide text attained the higher precision video retrieval than automatically recovered the spoken text.

Haojin Yang *et al.* (Smith, 2007) present an approach for automated lecture video indexing based on video OCR technology: Firstly, a video segmenter was developed for an automated slide video structure analysis. Secondly, the text detection was performed by localization and verification scheme. Here, SWT (stroke width transforms) was utilized for not only to remove false alarms from the text detection but also to analyze the slide structure further. To recognize the texts, a multi-hypothesis framework was adopted, which consists of multiple text segments, OCR, spell checking, and result merging processes. Finally, this algorithm was implemented for slide structure analysis and extraction by using the geometrical information of detected text lines. The accuracy of this approach was proven by evaluation.

3. MOTIVATION

Problem Statement

The major problem is to retrieve the content based lecture video from the input database based on the user query. Consider D be the input database, which consists of M number of videos. It is represented by $D = \{I_i;\ \ 1 \le i \le M\}$ which is given as input video for the proposed method. Subsequently, each video consists of a number of frames. The input video I contains K number of frames, which is defined as $I = \{F_p;\ \ 1 \le p \le K\}$. The major challenge is to retrieve the content based relevant video from the database D.

Challenges

- One of the most important challenges is how to retrieve the user's relevant video from this vast amount of information since each content or concept has a large number of videos (Beltran & Pla, 2016).
- The feature information being extracted from the frame is another challenge because the video includes some low-level information, as color, texture, shape, etc. (Yoo & Cho, 2007).
- The challenge (Cheny *et al.,* 2014) of using character and speech recognition method to extract the keywords from the frame based on the input query. The visual and textual keywords are presented in the frames of the video.
- The extraction of feature information is a challenge in the lecture video retrieval since it contains the homogenous composition and slide may also be obstructed by the person (Yang & Meinel, 2014).

4. PROPOSED METHODOLOGY: INTENT AWARE OPTIMIZATION USING GREY WOLF OPTIMIZER FOR LECTURE VIDEO RETRIEVAL

The ultimate aim of this paper is to retrieve the content based lecture video using the grey wolf optimizer. Normally, the large number of videos is available on the Internet. Depending on the input query, the relevant content based videos are retrieved. The proposed methodology contains the video retrieval and optimization algorithm. Initially, the key frame is extracted from the videos by the variance measure between frames. Then, the data or text feature for the input query is obtained by the Optical character recognition and Local vector pattern. After the feature is achieved, the PENN classifier is utilized for the classification purpose. The classifier yields the number of relevant videos in which the user selects one video for the further optimization. Consequently, we develop the novel intent aware optimization using the grey wolf optimizer to retrieve the appropriate video. The novel optimization is performed based on the clustering and selected video; the optimal centroid is determined by the grey wolf optimizer. Finally, the matching process is done between the selected video and clustering mechanism where the centroids are generated. Figure 1 shows the block diagram of the proposed methodology.

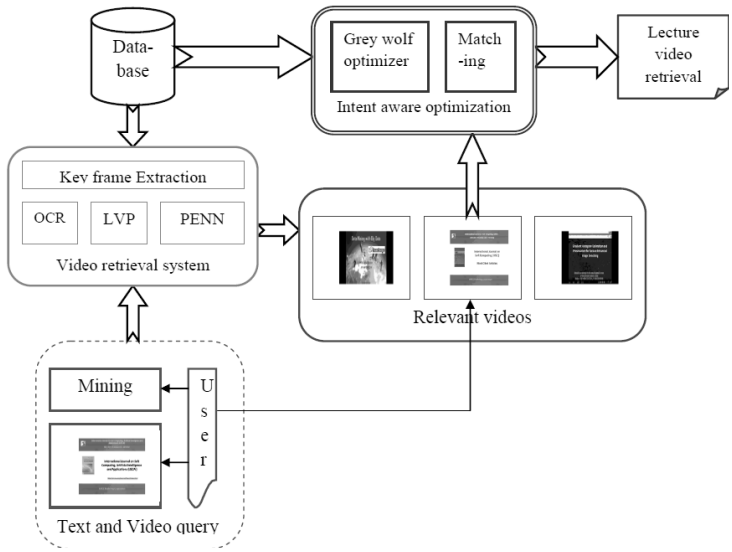


Fig. 1. Block diagram of proposed methodology.

Lecture Video Retrieval by classifier

The content-based lecture video is retrieved by the multimodal features and probability extended the nearest neighbor. Initially, the feature information of the video is extracted from the OCR and LVP based on the keyframes. Normally, the input video contains *F*-number of frames in which the keyframes are extorted for the feature information.

Keyframe extraction

The key frame extraction is the initial step for the content based lecture video retrieval. The input video contains a large number of frames from which we extract the exact frame, which is used for the subsequent steps. The keyframe is extracted by the variance measure between the frames. Higher difference value between the frame and its neighbor frame is considered as the key frame of the video. It is formulated as

$$d\left(F_j\right) = F_p - F_q, \quad p \neq q \tag{1}$$

$$F^k = \underset{j \in 1,2,...K}{\arg\max}\, d\left(F_j\right) \tag{2}$$

where j is the number of frames and F^k represents the resultant keyframe from the video. At last, m number of keyframes are extracted from the input video I.

Tesseract for OCR

After we extract the keyframes, the optical character recognition is applied to the keyframes for the feature information. Here, Tesseract (Smith, 2007; Smith *et al.,* 2009) is utilized in this paper for the feature extraction. Since the user gives the text or keyword to search the videos on the web, the character recognition is employed to extract the relevant text in the video. The OCR comprises line and word findings and word recognition. The feature information based on the text is apparently described below.

Text-Line and word findings

In this step, the text line in each frame is computed by the line finding and baseline fitting. On the other hand, the word finding is done through the fixed pitch detection, chopping, and the proportional word finding.

- The line finding includes two important keywords, blob filtering and line construction. The median height approximations evaluate the size of text and blobs are filtered out significantly. The filtered blobs are assigned in an appropriate line using the least median of squares fit. The quadratic spline is utilized here for the baseline fitting, which is used to resolve the constraints like an artifact in scanning and binding.
- After the text lines, the pitch text is estimated by the fixed pitch detection. Once the pitch text is detected, the words are separated by each character with the aid of chopping operation. Due to different font style and distinct line spacing between the two lines in the frame, the word findings become crucial. The OCR exploits the proportional word finding method.

Word recognition

Since the recognized word is not enough to retrieve the video, this step consists of segmentation and shape classification. Firstly, the blob sequence is undergone for the chopping operation. Then, the segmentation and search are performed to determine the candidate chop points. Secondly, the feature element is determined by the polygonal approximations, the features are trained by the classifier and recognize the word from each keyframe. The obtained feature is defined by

$$OCR\left(F^k\right)=G_h,\quad 1\le h\le w \tag{3}$$

where G represents extracted keyword and h is the number of the keyword. The feature from each key frame is also extracted by the Local Vector Pattern.

Local Vector Pattern

The texture descriptor of LVP (Fan & Hung, 2014) is utilized to extract the feature information from each keyframe. The LVP descriptor is applied to the key frame of the input video. The idea behind this descriptor is to generate the micro-patterns through pairwise directions of the vector with respect to its reference pixel and neighborhood pixel. The LVP is expressed as below

$$LVP_l\left(b,\phi\right)=\sum_{a=1}^{m}2^{a-1}h_l\left(a,b,\phi\right) \tag{4}$$

where $LVP_l(\bullet)$ represents the LVP at neighborhood distance l, b is the reference pixel and ϕ denotes the index angle and h is the histogram content of the keyframe.

$$h_l\left(a,b,\phi\right)=\begin{cases}1, & if \quad H^l_{(a,b,\phi)}\ge 0\\ 0, & Otherwise\end{cases} \tag{5}$$

where

$$H^l_{(a,b,\phi)}=P^l_{\phi+45}\left(a\right)-\left(\frac{P^l_{\phi+45}\left(b\right)}{P^l_{\phi}\left(b\right)}\times P^l_{\phi}\left(a\right)\right) \tag{6}$$

$$P^l_{\phi}\left(a\right)=F^k\left(\phi,l\right)-F^k\left(b\right) \tag{7}$$

where $F^k(\phi,l)$ is the intensity of the pixel with respect to the l distance and index angle ϕ. The feature vector is computed by the histogram of the texture image. The feature elements are determined by the following formula:

$$LVP\left(F^k\right)=\left\{C_q;\quad 1\le q\le 255\right\} \tag{8}$$

where C is the number of bins in the histogram image, and q ranges from 0 to 255. Finally, the feature of each video is evaluated by the concatenation of OCR and LVP pattern of the keyframes. It is expressed by

$$f(I) = \left\{OCR\left(F^k\right), LVP\left(F^k\right)\right\} \tag{9}$$

where I is the input video and F^k represents the keyframe. The computed feature vector is fed into the classifier to retrieve the relevant videos.

PENN classifier

Once the features are obtained, the PENN classifier is utilized to retrieve the videos based on the query. This classifier is employed to classify the desired videos based on the nearest neighbor. The advantage of using PENN classifier (Sanjay B. Waykar and C. R. Bharathi, 2017) over K-nearest neighbor is that the deficiency problem gets resolved. The classifier is performed by the two levels of neighbors of the video. Also, to compute the degree of membership, the PENN classifier is utilized with the aid of probability. Consider Q be the input query, which is given by the user that is either text query or video query. The text of each frame is computed from the optical character recognition. The similarity measure is evaluated between the input query and feature vector.

The similarity measure is calculated by the minimum value between the input query and feature. Then, the similarity measure between two keyframes is computed by the distance of the keyword and local vector pattern. It is represented as

$$S(Q, f_i) = \underset{x \in 1,2,..n}{\arg\min}\, SF^k\left(F_Q^k(x), F_{f_i}^k(x)\right) \tag{10}$$

where SF^k is the similarity measure of a keyframe, $F_Q^k(x)$ and $F_{f_i}^k(x)$ represent the query and feature of the x^{th} keyframe.

$$SF^k\left(F_Q^k(x), F_{f_i}^k(x)\right) = d_1\left[OCR\left(F_Q^k(x)\right), OCR\left(F_{f_i}^k(x)\right)\right] + d_2\left[LVP\left(F_Q^k(x)\right), LVP\left(F_{f_i}^k(x)\right)\right] \tag{11}$$

$$d_1(A,B) = \frac{1}{255}\sqrt{\sum_{c=1}^{255}(A_c - B_c)^2} \quad and \quad d_2(A,B) = 1 - \frac{(A \cap B)}{(A \cup B)} \tag{12}$$

After the queries are matched with the feature using the similarity measure, the N neighbor video of its corresponding query is selected by the probability function.

$$P\left(f_i^Q\right) = \frac{S(Q, f_i)}{\sum_{i=1}^{M} S(Q, f_i)} \tag{13}$$

where P is the probability measurement, and M is the number of videos in the input database. These neighbor videos are acted as video query to determine the corresponding relevant videos. The probability measure is calculated as

$$P\left(f_i^{f_i}\right) = \frac{S(f_i, f_i)}{\sum_{i=1}^{F} S(f_i, f_i)} \tag{14}$$

Finally, the cumulative probability is estimated by the weighting constant of the two probability measures. This function is used to determine the relevant videos of the input query which is given by

$$TP\left(f_i^Q\right) = \eta P\left(f_i^Q\right) + \lambda P\left(f_i^{f_i}\right) \tag{15}$$

where η and λ are the two weighting constants. The first term represents the probability measure of the input query. On the other hand, the second term denotes the probability measure of the neighbors of obtained relevant videos. Finally, the N relevant videos are retrieved from the database based on the input query.

Designing of Intent aware optimization using grey wolf optimizer and matching

Once we determine the relevant video, the user has to select one relevant video to retrieve the appropriate video for the input query. After that, the N number of relevant videos is obtained by the video retrieval system in which the user selects one video. The novel intent aware optimization is proposed in this paper for the lecture video retrieval. The optimization algorithm is designed to determine the appropriate video based on the matching process.

Step 1: User selection of video based on the feedback

Based on the input query, the N number of relevant videos is retrieved from the input database using PENN classifier. However, the user should not be aware which video is more relevant to the text or video query. To achieve this objective, the user has to select one video from the resultant relevant videos. The user selected video is represented as N'. This video is undergone for the matching process to retrieve the appropriate video for the user query. Simultaneously, the grey wolf optimizer is applied to the input database to determine the optimal solution.

Step 2: Grey wolf optimization algorithm

The grey wolf optimizer (Mirjalili *et al.,* 2014) is the recent optimization algorithm, which is inspired by the grey wolves. It is used to determine the optimal solution significantly by the hierarchy and hunting mechanism of the grey wolves. The hierarchy poses with four levels, which are alpha, beta, delta, and omega. Here, the alpha wolf is considered as the best search agent; beta and delta are defined as the second and third best search agents. The omega wolf follows the hunting behavior of these three wolves. The hunting phase includes tracking, chasing, encircling, and attack towards the solution. The hunting behavior has been mathematically modeled to solve the optimization problems. The algorithm is described below.

Algorithm Elucidation

To mathematically model the algorithm, we exploit the fittest solution as an alpha wolf, and the best solution is defined as beta and delta wolves. Here, the grey wolf optimization is used to determine the optimal solution.

i) Initially, e number of centroids are generated from the input database which includes M number of videos. It is represented as

$$Y = \{c_1, c_2, ..., c_e\} \tag{16}$$

where e is the number of centroids, which is optimized iteratively to find the best solution by the hunting mechanism.

ii) During hunting mechanism (Mirjalili *et al.,* 2014), the grey wolf has the tendency to identify the location of prey and encircle them. Usually, the hunt mechanism is guided by the alpha, beta, and delta wolves. To calculate the optimal solution, the position is updated by the three wolves. The mathematical formulation is expressed below.

$$Y(t+1) = \frac{Y_1 + Y_2 + Y_3}{3} \tag{17}$$

where

$$Y_1 = Y_\alpha(t) - U_1 \cdot P_\alpha, \quad Y_2 = Y_\beta(t) - U_2 \cdot P_\beta, \quad Y_3 = Y_\delta(t) - U_3 \cdot P_\delta \tag{18}$$

where Y_1, Y_2, Y_3 are the solution obtained by the grey wolf, $Y_\alpha, Y_\beta, Y_\delta$ represent the alpha, beta, and delta wolves, and P is the position vector. The position is determined by the current position of the solution and current position of the grey wolf. Then, the value z is decreased, which leads to update the position of the alpha, beta, and delta wolves. It is given by

$$P_\alpha = \left|V_1 \cdot Y_\alpha(t) - Y(t)\right|, \quad P_\beta = \left|V_2 \cdot Y_\beta(t) - Y(t)\right|, \quad P_\delta = \left|V_3 \cdot Y_\delta(t) - Y(t)\right| \tag{19}$$

where t represents the current iteration, and U and V are the coefficient vectors. Then, the position is updated for the subsequent iteration by the coefficient vectors and z. It is defined by

$$\begin{aligned} U &= 2z \cdot r_1 - z \\ V &= 2 \cdot r_2 \end{aligned} \tag{20}$$

where r_1 and r_2 are the random vectors, which range from zero to one and z is the value, which decreased from 2 to 0 over the iteration. The value is reduced iteratively, and the grey wolf updates their position to find the optimal solution of the centroid for the lecture video retrieval. The deduction of z value leads to the decrease of the U and V coefficient vectors for every iteration. The fitness value is calculated, which is used to update the position of the grey wolves.

iii) The fitness function is an important aspect of the optimization algorithm. The fitness value is used to determine the relevant videos from the input database based on the input query. The video is represented by $K \times 256$, where K is the number of frames and 256 features. Similarly, the centroid is also defined as $K \times 256$ which is generated by the grouping of certain videos. The fitness function is calculated by

$$fitness = \sum_{i=1}^{M} \min_{j \in (1,2,\ldots,e)} c_j(i) \tag{21}$$

where M is the number of videos in the input database, and e is the number of centroids.

Step 3: Attains appropriate relevant video by matching

Once we evaluate the fitness value, the matching process is undergone for the user query. The matching is done between the user selected video and the solution where the optimal centroids are obtained. The matching is undergone by the equivalent distance between the user selected videos and optimal videos obtained from grey wolf optimizer. The user selected video is represented as N'. It is expressed by

$$R_v = \sqrt{\sum_{i=1}^{e} \left(N'(K) - I_i'(K)\right)^2} \tag{22}$$

where $N'(K)$ is the user selected video, which contains K number of frames, $I_i'(K)$ represents the resultant videos using grey wolf optimizer, and e is the number of the centroids. If the user selected video matches with the second centroid, then the videos in the database, which have the minimum value of the second centroid, are taken as the appropriate videos for the input query. Thus, the lecture video is retrieved efficiently using the proposed intent aware optimization algorithm.

Table 1. Pseudocode of the Grey wolf optimization algorithm.

1	**Initialization**
2	Grey wolf population, $Y_i, i = 1,2,..,e$
3	z, Coefficient vectors U and V
4	**Begin**
5	Calculate the fitness of each search agent
6	Y_α = the best search agent
7	Y_β = the second best agent
8	Y_δ = the third best agent
9	**While** (t < max. number of iteration)
10	**For** each search agent
11	Update the position by $Y(t+1) = \frac{Y_1 + Y_2 + Y_3}{3}$
12	**End** for
13	Update z, U and V
14	Calculate the fitness of all search agent
15	Update Y_α, Y_β and Y_δ
16	t=t+1
17	**End** while
18	Return Y_α
19	**End**

5. RESULTS AND DISCUSSION

This section presents the experimental results for the proposed intent aware optimization for the lecture video retrieval. The performance is analyzed by the evaluation parameters and it is compared with the existing systems.

Experimental setup

Dataset description:

The video utilized for our experimentation is obtained from the publicly available resources (YouTube). Here, the 60 numbers of videos are used, categorized into different domains, such as data mining, image processing, soft computing, wireless communication, optical communication, and networking.

Evaluation metrics:

The performance of the proposed lecture video retrieval system using grey wolf optimizer is analyzed using the F-measure, recall, and precision. The performance is compared with the existing systems like KNN, ENN, and PENN for both text and video query.

a) In lecture video retrieval, the precision and recall are the two evaluation metrics, which are defined in terms of

a set of relevant and retrieved number of videos. The precision and recall are formulated as

$$Precision = \frac{R_{rl} \cap R_{rt}}{R_{rt}} \tag{23}$$

$$Recall = \frac{R_{rl} \cap R_{rt}}{R_{rl}} \tag{24}$$

where R_{rl} represents the number of relevant videos and R_{rt} is the number of retrieved videos.

b) F-measure is defined as the measure of accuracy, which is used to compute the value using precision and recall. It is defined by

$$F - measure = \frac{2 * precision * recall}{precision + recall} \tag{25}$$

Parameters fixed for analysis:

The value *k* and radius are the two parameters, which is used to analyze the performance of the proposed lecture video retrieval system. The *k* is a user-defined parameter since it represents the number of retrieved videos, which is given by the user. The radius is defined in the Local vector pattern and determines the best value. Finally, the performance is analyzed for the input video query based on the *k* and radius parameter.

Algorithms selected for comparison

KNN: The K-nearest neighbor algorithm is used mainly for the classification purpose. The feature data is given as input to the KNN (Wu *et al.,* 2007) algorithm. In the training data samples, this algorithm is used to group the k number of objects, which is close to the testing data samples. Here, the user gives the input as the video query for the lecture video retrieval, which is described in KNN+VQ. Then, the k nearest neighbors of videos are grouped, which is further used for the testing phase. Similarly, the text query is given as input for the lecture video retrieval using a k-nearest neighbor algorithm, which is explained in KNN+TQ.

ENN: The extended nearest neighbor (Tang & He, 2015) algorithm is used to solve the two-class classification problem. The nearest neighbor of test samples is considered for the classification. The ENN algorithm is easy for implementation and provides competitive classification. Here, the video query is considered as the input for the ENN+VQ. Simultaneously, the ENN+TQ algorithm describes the classification using extended nearest neighbor using the input as text query.

PENN: The Probability extended the nearest neighbor, which is the extension of the ENN classification algorithm. The PENN classifies features by the two levels of neighbors. The first one is defined by the probability measure between the query and keyframes. On the other hand, the probability measure is determined between the two frames of the video. The video query given to the PENN classification is deliberated in PENN+VQ. Consequently, the input is taken as the text query, which is given by the user to the PENN algorithm in PENN+TQ.

GWO: The proposed optimization algorithm is used to retrieve the lecture video by the grey wolf optimizer and input as video query. Initially, in the GWO+VQ, the relevant videos are retrieved by the video retrieval system. This system consists of key frame extraction, OCR, and LVP for feature extraction and PENN classification. After the videos are retrieved, the user has to select one video for the matching purpose and the grey wolf optimization (Mirjalili *et al.,* 2014) is applied to the input database simultaneously. This algorithm is used to determine the optimal solution of the centroids where the matching is done by the user selected video. After that, the text query is given by the user for the video retrieval using grey wolf optimizer, which is explained in GWO+TQ.

Experimental results

Figure 2 shows the experimental results of the lecture video retrieval. Here, the text query and video query are given as the input to the video retrieval system. Figure 2.a shows the input query as the video query. The proposed system retrieves the relevant videos based on the video query, which is represented in Figure 2.b. Similarly, Figure 2.c depicts the text query 'network', which is given as the input to the intent aware optimization of lecture video retrieval. Figure 2.d shows the acquired retrieved videos, which are relevant to the input text query. Figure 2.e depicts the GUI file of the proposed lecture video retrieval system. In this Figure 2.e, we can select the database and the input is given by the user, which is either text or video. From the relevant videos, the user selects one video, which is further used for the optimization by the grey wolf optimizer. Based on the matching between the videos, the final retrieval of the video is achieved.

Fig. 2. Experimental results.

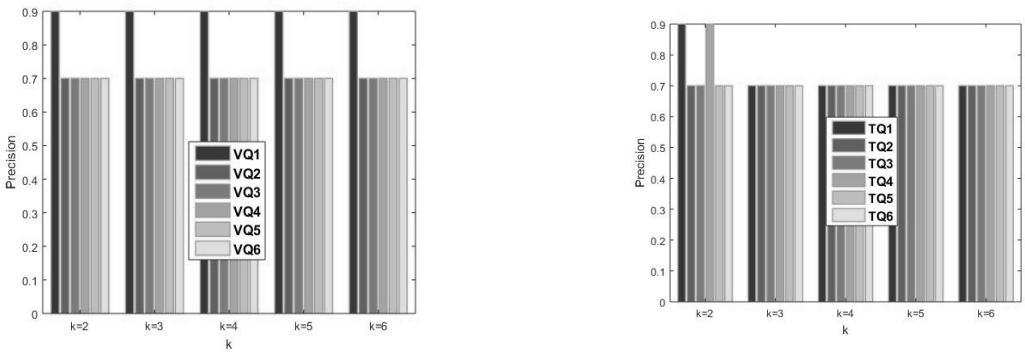
Performance Analysis

This section presents the performance analysis of the content based lecture video retrieval using grey wolf optimizer. The performance is analyzed by the evaluation metrics, such as F-measure, recall, and precision and the performance is compared with the existing systems like KNN+VQ, ENN+VQ, PENN+VQ, KNN+TQ, ENN+TQ, and PENN+TQ.

Analysis by k-value

Figure 3 depicts the precision performance analysis by the k-value. The precision is defined in terms of accuracy. It is a fraction measure of the retrieved videos that are relevant, which is also termed as positive predictive value. Figure 3.a shows the precision analysis for the input video query. The k value is represented by the number of retrieved videos the user needs. When the user wants four numbers of retrieved videos, the precision value for the first video query is 90%, whereas, regarding the precision value for all other input video queries, the value of 70% is achieved. Similarly, for the first video query, the high precision value of 90% is attained for all the k values, which are deliberated in Figure 3.a. Subsequently, the text query is given as the input for the proposed system, and its performance is shown in Figure

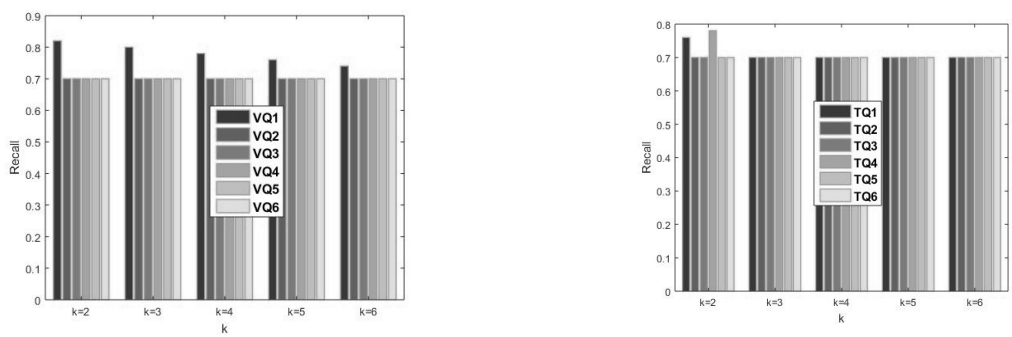
3.b. The TQ4 attains the precision value of 90% and 70% based on the user requirement. When the text query TQ5 is given as input to the proposed system, the precision value is 70% for all the k values, which is demonstrated in the Figure 3.b. The high precision value for both text and video query ensures the better retrieval performance.



(a) Video query (b) Text query

Fig. 3. Performance analysis by precision.

Figure 4 shows the performance analysis using the recall metric. The recall is also defined as the sensitivity, which is the measure of relevant video that is retrieved. Figure 4.a depicts the recall performance analysis based on the k value. When using the first video query as input, also based on the user needs, the recall value of 82% for k=2, 80% for k=3, attains 78% when k=4, 76% at k=5 and finally 74% of recall is achieved if k=6, which is represented in Figure 4.a. The rest of the input video queries attain the 70% value for all the k values. Simultaneously, Figure 4.b shows the recall performance for the text query. When the user requirement is two, the 76% of recall value is obtained for the TQ1, and 78% value for the TQ4 and also 70% of recall are achieved for the other input text queries. Similarly, when the user needs five relevant videos based on the text query TQ, the proposed system acquires 70% of recall value, which is demonstrated in Figure 4.b.

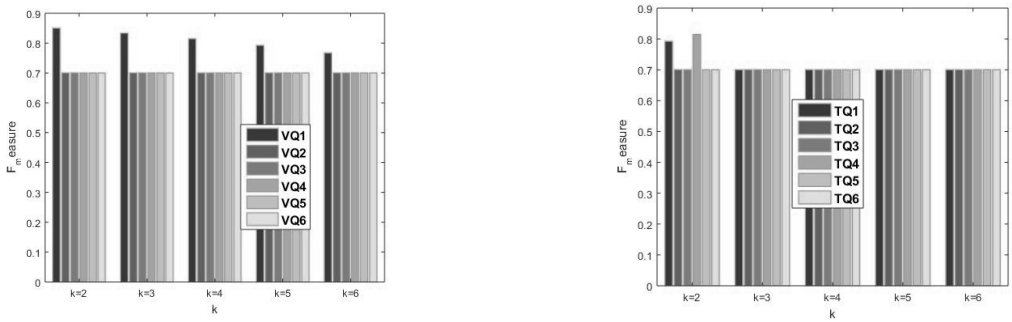


(a) Video query (b) Text query

Fig. 4. Performance analysis by recall.

The performance analysis for the F-measure is shown in Figure 5. The F-measure is one of the evaluation metrics, which is defined by the fraction of precision and recall value. The video and text query are given as input to the proposed intent aware optimization. Figure 5.a represents the trade-off between the f-measure and k value. When the k value is one, the first video query attains the 85% value, which is gradually decreased to 76.6%. Rather than the other video queries, the VQ1 attains the better f-measure performance, which is demonstrated in Figure 5.a. The text query for the f-measure performance is analyzed, shown in Figure 5.b. While using the input as TQ3, the f-measure value

of 70% obtained depends on the user requirement. On the other hand, while using TQ4, the higher value of 81.4% f-measure is obtained. When the k value is two, 79.2% of f-measure is attained for the first text query. Figure 5 shows the higher f-measure value of 85% for video query and 81.4 for text query.

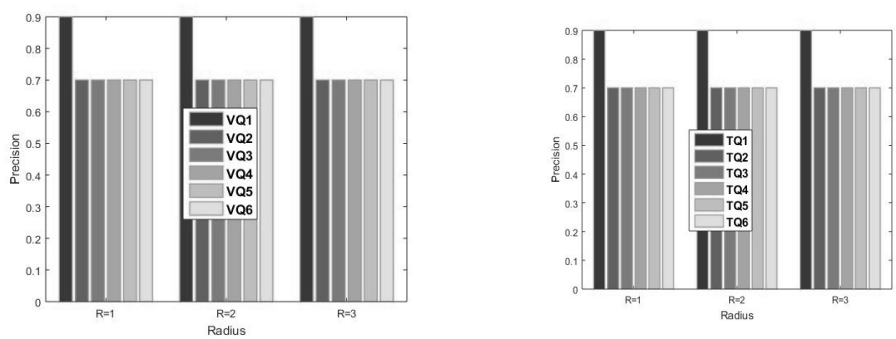


(a) Video query (b) Text query

Fig. 5. Performance analysis by F-measure.

Analysis using radius

This section presents the performance analysis of both text and video query using the parametric like precision, recall, and F-measure. Figure 6 depicts the performance analysis by the precision parameter. Figure 6.a shows the precision performance analysis based on the user video query. The radius is the value that is fixed in the local vector pattern, which determines the texture feature information from the input video. While using the VQ1 as the input query for the proposed system, the higher precision of 90% is obtained based on the radius. Similarly, the 70% precision value is achieved for the other video queries, which are represented in Figure 6.a. Subsequently, Figure 6.b. shows the precision performance of the proposed method while using the input as text query. When the radius is three, the text query TQ1 attains the 90% of precision and 70% of precision for the TQ2 to TQ6. While using the text query also, the higher precision value of 90% is obtained based on the radius value, which is demonstrated in Figure 6.b.

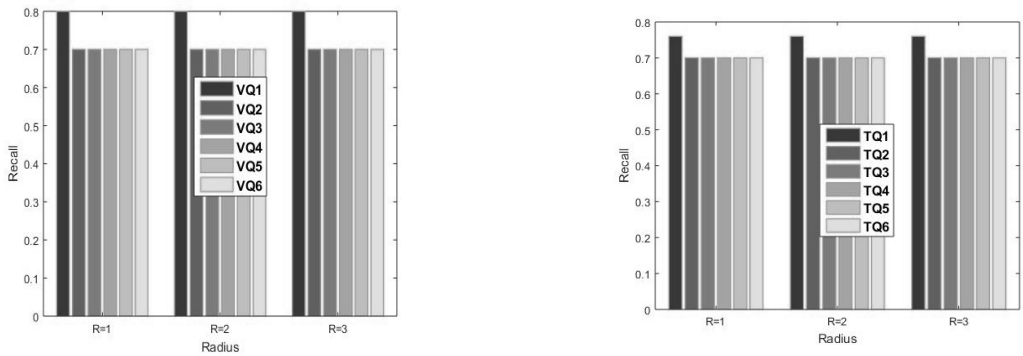


(a) Video query (b) Text query

Fig. 6. Performance analysis by precision.

Figure 7 represents the recall performance analysis for the text and video query. The recall is the measure that is defined by the ratio of a number of retrieved relevant video to the total number of relevant videos. Figure 7.a shows the performance analysis for the video query. While giving the input as VQ5, the recall value of 70% is attained based on the radius value. When the radius in local vector pattern is one, the higher recall value of 80% is obtained by the input video query VQ1 and 70% value for other video queries, which is deliberated in Figure 7.a. Simultaneously,

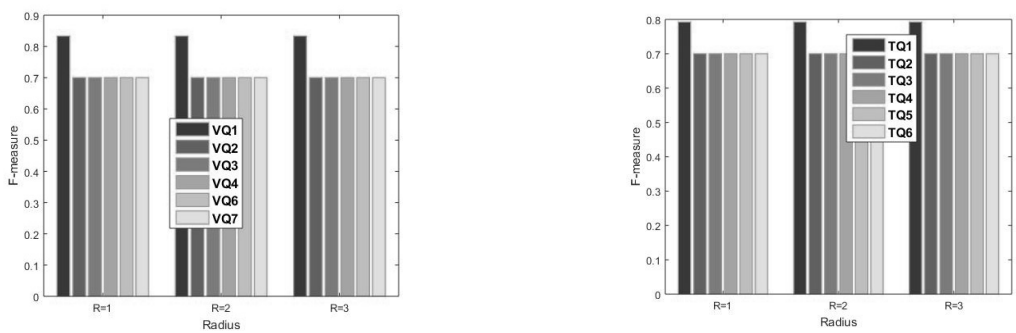
the text query is given as input to the proposed intent aware optimization algorithm, and its performance is analyzed in Figure 7.b. The 76% of higher recall value is achieved for the text query TQ1. 70% value is obtained for all the video queries based on the radius value. While using radius of two, the TQ1 acquires the 76% of recall value, which is represented in Figure 7.b. From the figure, it can be shown that the proposed system has the better performance of the lecture video retrieval.



(a) Video query (b) Text query

Fig. 7. Performance analysis by recall.

The F-measure performance analysis for the proposed system is shown in Figure 8. The F-measure is used to evaluate the value using precision and recall parameter. Figure 8.a depicts the performance analysis for the text video query. Based on the radius in LVP, the F-measure gets varied for the various input video query. While using the video query VQ2, the proposed system achieves 70% of f-measure value based on the fixed parameter radius. Similarly, when the LVP uses R=3, the higher value of 83.3% of f-measure is achieved by the query VQ1, which is deliberated in Figure 8.a. Figure 8.b shows the performance analysis for the text query. The user gives the text as input to the lecture video retrieval. While using the text query TQ6, the f-measure value of 70% is obtained for the radius value R=1, R=2, and R=3. But the better performance for the text query is ensured by the f-measure value of 79.2%, which is shown in Figure 8.b.



(a) Video query (b) Text query

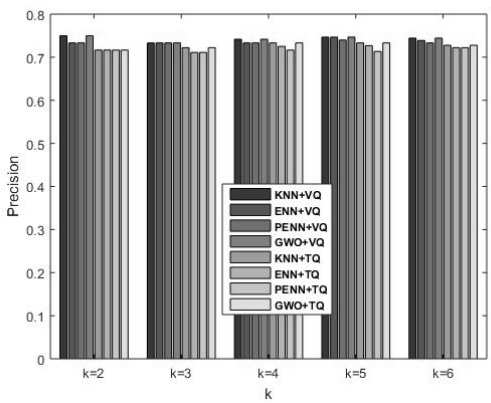
Fig. 8. Performance analysis by F-measure.

Comparative analysis

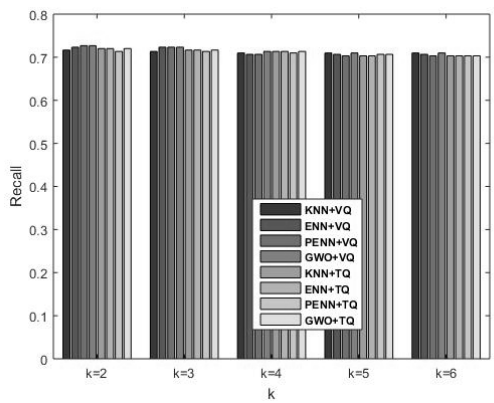
This section presents the comparative analysis of the proposed method with the existing systems, like KNN+VQ, ENN+VQ, PENN+VQ, KNN+TQ, ENN+TQ, and PENN+TQ. Figure 9 represents the comparative performance, which is analyzed by the precision, recall, and f-measure. Figure 9.a shows the comparative performance analysis

by the precision. The precision is defined in terms of accuracy value, which is expressed by percentage. When the user requirement is four numbers of retrieved videos, the existing system such as KNN+VQ system attains 74.1%, ENN+VQ achieves 73.33%, KNN+TQ acquires 73.3%, and ENN+TQ obtains 72.5%. Similarly, the other existing PENN classifier attains 73.3% for video query and 71.6% for text query. However, compared to the existing systems, the proposed system achieves 74.17% for video query and 73.33% for text query, which is shown in Figure 9.a

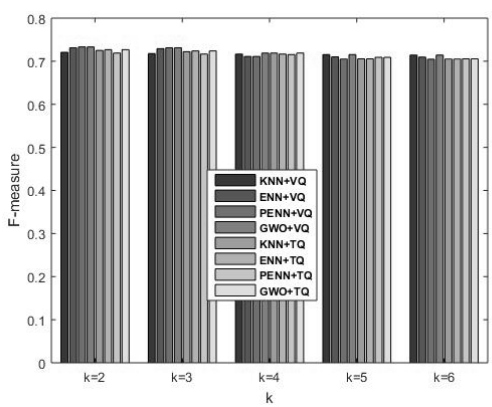
Figure 9.b depicts the comparative performance analysis of the recall. The existing PENN classifier for video query achieves 72.6%, 72.3%, 70.6%, and 70.3% for the different k value. But, the GWO+VQ achieves the higher recall value of 72.67% for the video query and also obtains 72% for the text query when compared to the existing systems, which is represented in the Figure 9.b. Consequently, Figure 9.c depicts the comparative performance analysis by the F-measure. The F-measure is used to determine the value by both precision and recall value. When the k value is five, the existing KNN algorithm attains 71.5% for VQ and 70.5% for text query, the ENN+VQ achieves 71.03% and 70.56% for ENN+TQ, and PENN classifier acquires 70.5% and 70.8% for video and text query. The proposed intent aware optimization attains a higher value of 73.3% for video query and 72.67% for text query when compared to the KNN+VQ, ENN+VQ, PENN+VQ, KNN+TQ, ENN+TQ, and PENN+TQ, which is shown in Figure 9.c. The proposed system ensures the better performance for the content based lecture video retrieval.



(a) Precision



(b) Recall



(c) F-measure

Fig. 9. Comparative performance analysis.

Comparative Discussion

This section presents the comparative discussion of the proposed method and the existing methods, like KNN+VQ, ENN+VQ, PENN+VQ, KNN+TQ, ENN+TQ, and PENN+TQ. The existing methods, such as KNN+VQ, ENN+VQ, PENN+VQ, KNN+TQ, ENN+TQ, and PENN+TQ, attain the precision of 0.7500, 0.7467, 0.7400, 0.7333, 0.7267, and 0.7222, while the proposed method attains the precision of 0.7500 for video query and 0.7333 for text query.

The proposed method has the recall of 0.7267 for video query and 0.7200 for text query; on the other hand, the existing methods attain the recall of 0.7167, 0.7233, 0.7267, 0.7200, 0.7200, and 0.7133, respectively. Similarly, the proposed method has the maximum F-Measure than the existing methods for video query and text query. The computation time of the proposed method is 5.5 seconds for video query and 5.2 seconds for text query. On the other hand, the existing methods, like KNN+VQ, ENN+VQ, PENN+VQ, KNN+TQ, ENN+TQ, and PENN+TQ, have the computation time of 7, 8, 6, 9, 8.5, and 6.5 seconds, respectively.

Table 2. The comparative discussion of the proposed method with the existing methods.

	Precision	**Recall**	**F-Measure**	**Computation Time (Sec)**
KNN+VQ	0.7500	0.7167	0.7208	7
ENN+VQ	0.7467	0.7233	0.7311	8
PENN+VQ	0.7400	0.7267	0.7333	6
KNN+TQ	0.7333	0.7200	0.7250	9
ENN+TQ	0.7267	0.7200	0.7267	8.5
PENN+TQ	0.7222	0.7133	0.7190	6.5
GWO+VQ	0.7500	0.7267	0.7333	5.5
GWO+TQ	0.7333	0.7200	0.7267	5.2

6. CONCLUSION

In this paper, we presented the intent aware optimization for the content based lecture video retrieval using the grey wolf optimizer. Normally, each video consists of a number of frames. Initially, in the proposed methodology, the keyframe was extracted from each video by the variance measure. After the key frame is extracted, the OCR and LVP patterns were applied to extract the feature information. The keywords were obtained based on the user input query. The feature vector was then fed into the PENN classifier where the two levels of neighbors were measured. The PENN classifier was used to retrieve the relevant videos for the input query by the probability measurement between the frames. Then, the user had to select one relevant video for the further optimization. Consequently, the new intent aware optimization was developed by GWO to retrieve the appropriate lecture video. The grey wolf optimizer was utilized to determine the optimal solution of the centroid. The GWO was applied to the input database where the matching was done by the user selected video. After matching was done, the appropriate content based lecture video was retrieved. The outcome of the proposed intent aware optimization achieved higher precision of 75% value, which proved the better video retrieval performance.

REFERENCES

Aly, R., Doherty, A., Hiemstra, D., Jong, F.D. & Smeaton, A.F 2013. The uncertain representation ranking framework for concept-based video retrieval, *Information retrieval,* 16(5): 557-583.

Balasubramanian, V., Doraisamy, S.G & Kanakarajan, N.K. 2015. A multimodal approach for extracting content descriptive metadata from lecture videos, *Journal of Intelligent Information Systems,* 46(1): 121-145.

Beltran, R.F & Pla, F. 2016. Latent topics-based relevance feedback for video retrieval, *Pattern recognition*, (51): 72-84.

Chen, H., Cooper, M., Joshi, D. & Giro, B. 2014. Multi-modal Language Models for Lecture Video Retrieval, *in Proceedings of the ACM International Conference on Multimedia, 1081*-1084.

Cheny, H., Cooperz, M., Joshiz, D & Girod, B. 2014. Multi-modal Language Models for Lecture Video Retrieval, *ACM International Conference on Multimedia*, 1081-1084.

Cooper, M. 2013. Presentation Video Retrieval using Automatically Recovered Slide and Spoken Text, *Multimedia content and Mobile devices,* 8667: 1-7.

Fan, K.C & Hung, T.Y 2014. A Novel Local Pattern Descriptor—Local Vector Pattern in High-Order Derivative Space for Face Recognition, *IEEE transactions on image processing,* 23(7): 2877-2891.

Feng, B., Cao, J., Bao, X., Bao, L., Zhang, Y., Lin, S & Yun, X. 2011. Graph-based multi-space semantic correlation propagation for video retrieval, *The Visual Computer,* 27(1): 21-34.

Gao, X., Li, X., Feng, J & Tao, D. 2014. Shot-based video retrieval with optical flow tensor and HMMs, *Pattern recognition letters,* 30(2): 140-147.

Hernandez, M.C. & Hernandez, A.C. 2014. Content-Based Video Retrieval System for Mexican Culture Heritage based on Object Matching and Local-Global Descriptors, *In Proceedings of IEEE International Conference on Mechatronics, Electronics and Automotive Engineering* (ICMEAE), 38-43.

Li, K., Wang, J., Wang, H & Dai, Q. 2015. Structuring Lecture Videos by Automatic Projection Screen Localization and Analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 37(6): 1233-1246.

Liao, K., Liu, G., Xiao, L & Liu, C. 2013. A sample-based hierarchical adaptive K-means clustering method for large-scale video retrieval, *knowledge-based system,* 49: 123-133.

Mirjalili, S., Mirjalili, S.M. & Lewis, A. 2014. Grey Wolf Optimizer, *Advances in Engineering Software,* 69: 46-61.

Nguyen, N.V., Coustaty, M & Ogier, J.M. 2014. Multi-modal and cross-modal for lecture videos retrieval, *In Proceedings of IEEE International Conference on Pattern Recognition (ICPR),* 2667-2672.

Sanjay B. Waykar & C. R. Bharathi. 2017. Multimodal Features and Probability Extended Nearest Neighbor Classification for Content-Based Lecture Video Retrieval, *Journal of Intelligent Systems*, 26(3): 585-599.

Smith, R. 2007. An Overview of the Tesseract OCR Engine, *In Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR),* 2: 629-633.

Smith, R., Antonova, D & Lee., D. 2009. Adapting the Tesseract Open Source OCR Engine for Multilingual OCR, *in Proceedings of the International Workshop on Multilingual OCR*.

Tang, B. & He, H. 2015. Enn: Extended nearest neighbor Method for Pattern Recognition, *Computational Intelligence Magazine, IEEE*, 10(3):52 – 60.

Va, S.C. & Narayanana, N.K., 2015. Key-frame extraction by analysis of histograms of video frames using statistical methods, *International Conference on Eco-friendly Computing and Communication Systems,* 70: 36-40.

Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.H., Steinbach, M., Hand, D.J & Steinberg, D. 2007. Top 10 algorithms in data mining, *Knowledge Information System,* 14(1): 1–37. https://www.youtube.com.

Yang H & Meinel, C. 2014. Content-Based Lecture Video Retrieval Using Speech and Video Text Information, *IEEE Transactions on Learning Technologies,* 7(2): 142-154.

Yang, H., Siebert, M., Luhne, P. & Sack, H. 2011. Lecture Video Indexing and Analysis Using Video OCR Technology, *In Proceedings of IEEE International Conference on Signal-Image Technology and Internet-Based Systems,* 54-61.

Yoo, H.W & Cho, S.B 2007. Video scene retrieval with interactive genetic algorithm, *Multimedia Tools, and Applications,* 34(3): 317-336.

Submitted: 07/07/2017
Revised: 05/11/2017
Accepted: 11/01/2018

إيجاد الحل الأمثل لإدراك القصد من استرجاع فيديو عن محاضرة استناداً على المحتوى باستخدام خوارزمية Grey Wolf Optimizer (GWO)

***سانجاي وايكار وسي آر بهاراتي**

*قسم علوم وهندسة الكمبيوتر، معهد البحث والتطوير للعلوم والتكنولوجيا، أفادي، تشيناي، الهند
قسم الإلكترونيات والاتصالات، معهد البحث والتطوير للعلوم والتكنولوجيا، أفادي، تشيناي، الهند

الخلاصة

في الوقت الحاضر، يتم استخدام تسجيلات الفيديو على نطاق واسع وبسهولة لنشر المعرفة بين الطلاب. ونظراً للتطور السريع في تقنيات التسجيل والتعلم القائم على تسجيلات الفيديو، تم نشر عدد كبير من مقاطع الفيديو على الإنترنت. ويكمن التحدي الرئيسي في استرجاع الفيديو المناسب بناءً على متطلبات المُستخدم. يقترح هذا البحث ايجاد الحل الأمثل لإدراك القصد استناداً على خوارزمية GWO لاسترجاع فيديو عن محاضرة. الخطوة الأولى في النظام المقترح هي استخراج الإطار الرئيسي. والخطوة التالية هي استخراج الإطار الرئيسي حيث يتم التعرف على الكلمات الرئيسية منه بواسطة التعرف الضوئي على الحروف ونمط المتجه المحلي (LVP). وبعد استخراج الميزات، تم استخدام مصنف PENN لاسترداد مقاطع الفيديو ذات الصلة للاستعلام النصي أو المسجل عن طريق الفيديو. بعد ذلك، يختار المستخدم فيديو واحد بغرض التطابق استناداً على الحل الأمثل. تم تطبيق خوارزمية GWO على قاعدة بيانات الإدخال حيث تحقق مهمة التجميع الحل الأمثل. وأخيراً، تم مطابقة الفيديو الذي حدده المستخدم مع الحل الأمثل لاسترجاع الفيديو المطلوب الاستعلام عنه. تم التحقق من صحة النتائج التجريبية، وكانت المعلمات التي استخدمت لتحليل الأداء هي مقياس F، الاستدعاء والدقة. تم مقارنة الأداء مع الأنظمة الحالية باستخدام تطبيق MATLAB. وتم الحصول على أعلى قيمة دقة بنسبة ٪75 من خلال الطريقة المقترحة التي تضمن الاسترجاع الفعال لفيديو المحاضرة استناداً على المحتوى.